Novogene | Providing leading genomic services & solutions

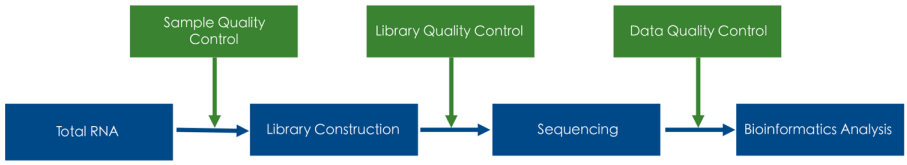
RNA-seq Analysis Report

Contract.Information	Contract.Content
Contract ID	H202SC22053202
Batch ID	X202SC22053202-Z01-F001
Species and Version	mm10
Project Name	Davis-US-UTHH-8-Mouse-EukmRNAseq-6Gb-WBI-Quantification-NVUS2022052356
Report Time	2022-07-27

1 Workflow of RNA sequencing projects

A transcriptome is a set of all the transcripts in one cell or one population of cells at certain status. Transcriptome analysis assists to study the identification of genes that are differentially expressed in distinct cell populations. Researchers can also gain a deeper insight into gene boundary identification, variable cleavage and transcript variation. [1]. RNA sequencing via Illumina platforms, based on the mechanism of SBS (sequencing by synthesis), offers a wide range of benefits on high throughput and high accuracy out of low sample requirements. This technical method can be a powerful tool for researching RNA transcriptional activity.

RNA sequencing projects are carried out as follows:



Novogene Co.,Ltd.

2 Analysis Pipeline

This is the workflow for medical species (human/mouse) mRNA sequencing data of standard bioinformatic analysis with a well-annotated reference genome, as follows:

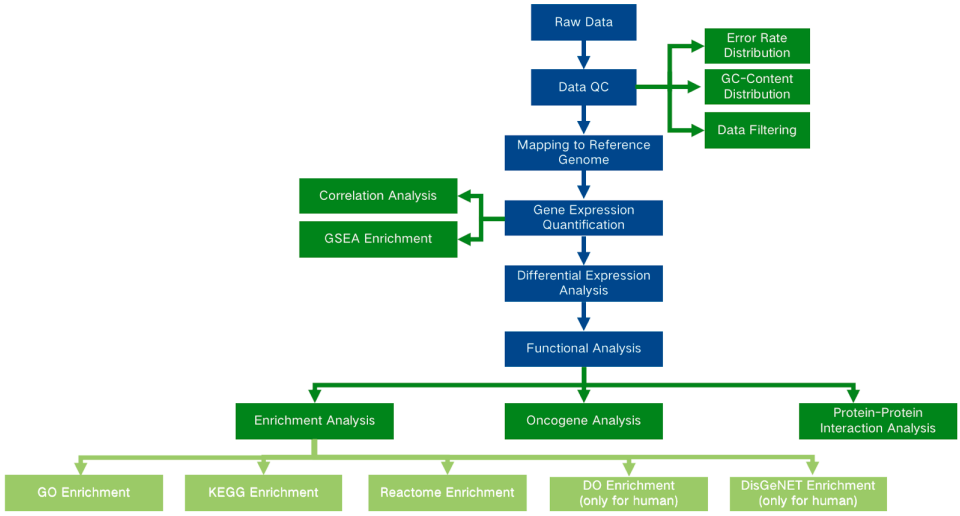


Figure 2.1 RNA-seq information analysis technology flow

For the analysis content of the above figure, if it exists in the content of the contract information analysis, the analysis is performed; if it does not exist, it is not performed.

Novogene Co.,Ltd.

2.1 Project Results

2.1.1 Raw data

Original image data file from high-throughput sequencing platforms (like Illumina) is transformed to sequenced reads (called Raw Data or Raw Reads) by CASAVA base recognition (Base Calling). Raw data are stored in FASTQ(fq) format files which contain sequences of reads and corresponding base quality. Each read has four descriptive lines, as indicated below:

Each read has four descriptive lines, as indicated below:

```
@A01426:11:H73WJDSX2:1:1101:2329:1016
1:N:0:GTGAGATC+GAACTAGC
TCGCACGCGCTGCCGTATGTGACGCCGTCGCTGCCGCACACAGGATCGTAGAGGCT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFF:FFFFFFFFFF
```

1. Line 1: Begins with the at sign (@) followed by sequence identifiers and optional description (such as FASTA header).
2. Line 2: Base sequences (raw read, A, G, C, and T).

3. Line 3: Begins with the plus sign (+) optionally followed by the same Illumina sequence identifiers and description information as in Line 1.
4. Line 4: The quality values for each base, corresponding to the data in Line 2.

Illumina Sequence Identifier:

Identifier	Meaning
A01426	Instrument – unique identifier of the sequencer
11	run number – Run number on instrument
H73WJDSX2	Flowcell ID - ID of flowcell
1	LaneNumber - positive integer
1101	TileNumber - positive integer
2329	X - x coordinate of the spot. Integer which can be negative
1016	Y - y coordinate of the spot. Integer which can be negative
1	Read Number - 1 can be single read or Read 2 of paired-end
N	Y if the read is filtered out (did not pass), N otherwise
0	control number - 0 when none of the control bits are on, otherwise it is an even number
GTGAGATC+GAACTAGC	Illumina index sequences

The details of Illumina Sequencing identifier are as follows:

1. A01426:11 A01426, Instrument - unique identifier of the sequencer; 11, Run number - Run number on instrument
2. H73WJDSX2:1:1101:2329:1016 means the coordinate of read on H73WJDSX2 (Flowcell ID) flowcell, line 1, 1101 tile is(x=2329, y=1016)
3. 1:N:0:GTGAGATC+GAACTAGC the first number is 1 or 2, 1 refers to single reads or the first read of paired ends, 2 refers to the second of paired ends; the second letter means whether reads is adjusted(Y means yes, N means no); the third number represent the number of Control Bits in sequence; six bases on the fourth place is Illumina index sequence.

2.1.2 Data Quality Control

The error rate for each base can be transformed by the Phred score as in equation 1. "e" represents sequencing error rate, "Qphred" represents base quality values of Illumina platforms (equation 1: Qphred = -10log10(e)).

Table 2.1 Illumina Casava 1.8 version base recognition and Phred score Concise correspondence between them

Phred.score	Base.Calling.error.rate	Base.Calling.correct.rate	Q.sorce
10	1/10	90%	Q10
20	1/100	99%	Q20
30	1/1000	99.9%	Q30
40	1/10000	99.99%	Q40

Sequencing error rate and base quality varies depending on sequencers, reagent residues, and different sample types. For RNA-seq technology, sequencing error rate distribution can be featured as below:

- Error rate increases with the sequencing reads for consumption of sequencing reagent. It is common in the Illumina high-throughput sequencing platform (Erlich Y, Mitra PP et al.2008; Jiang L, Schlesinger F et al.2011).
- The first six bases have a relatively high error rate due to the incomplete binding of random hexamers used in priming cDNA synthesis (Jiang et al.2011). In general, a single base error rate should be lower than 1%.

The sample sequencing error rate distribution is shown in the results file QC/1.Error (./result_tree.html). We also provide PDF and svg two types of image files.

CKOMGFP

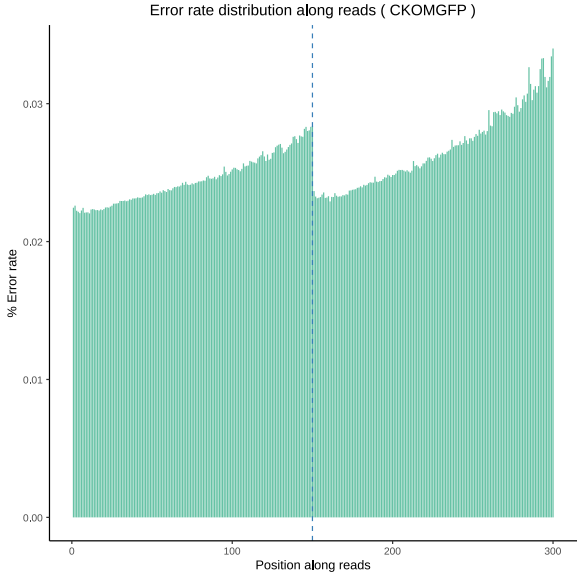


Figure 2.2 Sequencing data error rate distribution

The x-axis shows the base position along with each sequencing read and the y-axis shows the base error rate.

2.1.3 GC Content Distribution

GC content distribution indicates potential AT/GC separation which affects subsequent gene expression quantification. In view of random fragmentation and biological law of G/C-A/T content, G and C, A and T should be respectively equal, and the content should be stable throughout the entire sequencing process for the non-stranded library (If the library is strand-specific, AT separation or GC separation may occur). Large variations of sequencing error in the first 6-7 bases are allowed considering the use of random primer in library construction. It is normal that the first few bases have certain preferences in existing high-throughput sequencing technology. Result Directory: QC/2.GC (./result_tree.html)

CKOMGFP

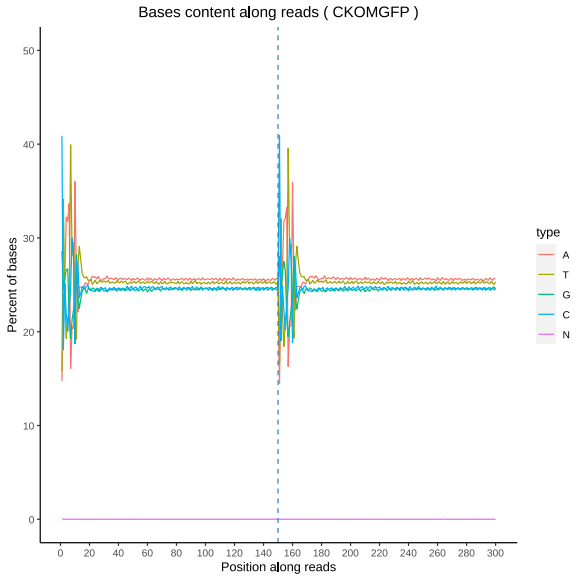


Figure 2.3 GC content distribution

The x-axis shows each base position within a read, and the y-axis shows the percentage of each base, with each base represented by a different color.

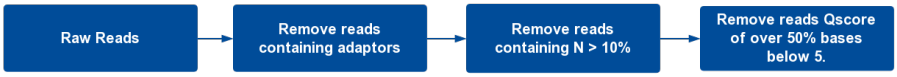
The left side of the vertical dashed line is the GC-content of read 1, the right side is the GC-content of read 2.

Novogene Co.,Ltd.

2.1.4 Data Filtering

The sequencing reads/raw reads often contain low-quality reads or reads with adapters, which will affect the quality of downstream analysis. To avoid this, it is necessary to filter the raw reads and obtain the clean reads.

Raw reads filtering is as follows:



- Remove reads with adapter contamination.
- Remove reads when uncertain nucleotides constitute more than 10 per cent of either read (N > 10%).
- Remove reads when low-quality nucleotides (Base Quality less than 5) constitute more than 50 per cent of the read.

Adapter Sequences:

Adapter	Sequence
P5 Adapter	P5-AATGATACGGCGACCACCGAGA (5’-3’)
	P5'-TTACTATGCCGCTGGTGGCTCT (3’-5’)
P7 Adapter	CGTATGCCGTCTTCTGCTTG-P7' (5’-3’)
	GCATACGGCAGAAGACGAAC-P7 (3’-5’)

For the sequencing data of each sample, see the results file: QC/3.Filter (./result_tree.html), as shown below.

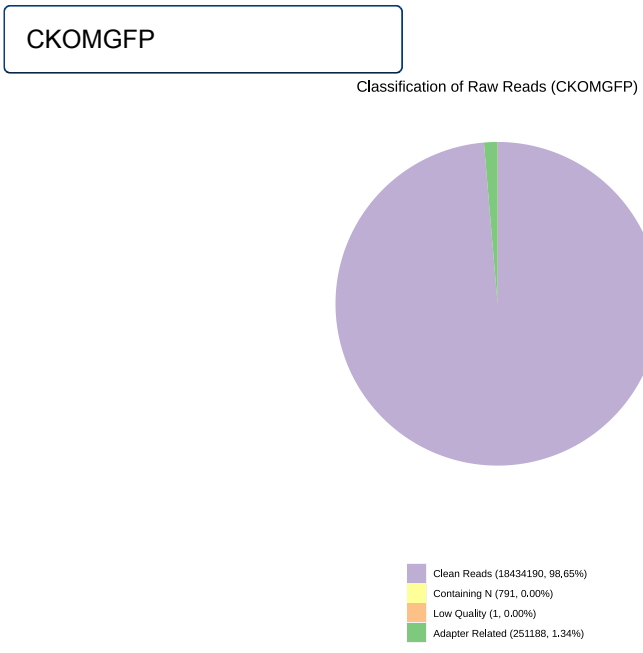


Figure 2.4 Sample Sequencing Data Filtering

Note: The proportions of the different colors in the graph represent the proportion of different components.

- **Adapter related**: (reads containing adapter) / (total raw reads).
- **Containing N**: (reads with more than 10% N) / (total raw reads).
- **Low quality**: (reads of low quality) / (total raw reads).
- **Clean reads**: (clean reads) / (total raw reads).

2.1.5 Data Quality Control Summary

Data are summarized in the table below.See the result file: QC/4.Stat/data_table.xls (./result_tree.html).

Table 2.2 Sample Sampling Data Quality Summary

sample	library	raw_reads	raw_bases	clean_
HKOSATGFP	CRRA220038496-1a	38286962	5.74G	3
HWTSATGFP	CRRA220038499-1a	23276328	3.49G	2
CKOMGFP	CRRA220038502-1a	37372340	5.61G	3
CKOSATGFP	CRRA220038503-1a	52473818	7.87G	5
CKOVATGFP	CRRA220038504-1a	42793526	6.42G	4
CWTMGFP	CRRA220038506-1a	49345140	7.4G	4
CWTSATGFP	CRRA220038507-1a	51499118	7.72G	5
CWTVATGFP	CRRA220038508-1a	30082250	4.51G	2

Showing 1 to 8 of 8 entries Previous 1 Next

- **sample**: SampleID.
- **library**: Library ID.
- **raw_reads**: Reads count from the raw data, four rows as a unit, with statistics of reads count for every sequencing.
- **raw_bases**: Base number of raw data. (number of raw reads) * (sequence length), converting unit to G.
- **clean_reads**: Base number of raw data after filtering. (number of clean reads) * (sequence length), converting unit to G.
- **clean_bases**: (clean base=clean reads*150bp) number multiply read length, saved in G unit.
- **error_rate**: Average sequencing error rate, which is calculated by

Qphred=-10log10(e).

- **Q20**: The percentage of the bases whose Q Phred values is greater than 20. (Number of bases with Q Phred value > 20) / (Number of total bases) *100.
- **Q30**: The percentage of the bases whose Q Phred values is greater than 30. (Number of bases with Q Phred value > 30) / (Number of total bases) *100.
- **GC_pct**: The percentage of G&C base numbers of total bases.(G&C base number) / (Total base number)*100.

2.1.6 Data Quality Control Q&A

Interpretation of Relevant Nouns

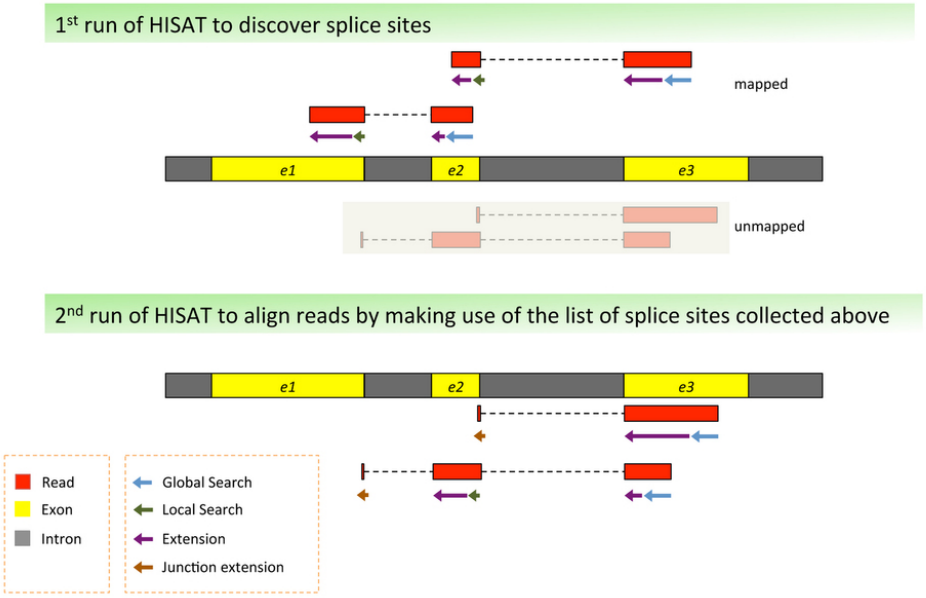
- **Q20**:The percentage of the total bases have Phred greater than 20. Phred=-10log10(e)
- **Q30**:The percentage of the total bases have Phred greater than 30. Phred=-10log10(e)

The sequencing error rate increases as the sequence's length increases. What is the acceptable error rate?

Novogene set a high standard for sequencing quality. Generally, the error rate of a single base should be lower than 1%. In some special cases, the maximum error rate of a single base should not be greater than 6%.

2.2 Alignment

Perform alignments with HISAT2 (http://ccb.jhu.edu/software/hisat2/faq.shtml) to the reference. HISAT2[4] uses a graph-based alignment and has succeeded HISAT and TOPHAT2. HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads. Besides one global GFM index, HISAT2 also includes a large set of small GFM indexes that collectively cover the whole genome. These small indexes (local indexes), combined with multiple alignment strategies, enabled effective alignment of RNA-seq reads, particularly, reads spanning multiple exons. The following figure shows the algorithm of split reads comparison by HISAT2:



The algorithm of hisat is divided into three parts:

1. The whole sequence was aligned to a single exon.
2. The sequencing sequence was piecewise aligned to two exons of the genome.
3. The sequencing sequence was segmented and aligned to more than three (including three) exons of the genome.

2.2.1 Mapping Result

The data are summarized in the table below. The sample and reference genome comparisons can be found in the results file: Mapping/2.Stat/align_pct.xls (./result_tree.html).

Table 2.3 Comparison of sample and reference genomes

sample	total_reads	total_map	unique_map	

HKOSATGFP	37437188	34003558(90.83%)	32610411(87.11%)	139
HWTSATGFP	22444206	20732562(92.37%)	19976817(89.01%)	75
CKOMGFP	36868380	35546076(96.41%)	34158474(92.65%)	138
CKOSATGFP	51447724	48305271(93.89%)	46507014(90.4%)	17
CKOVATGFP	41948516	40434850(96.39%)	39007736(92.99%)	14
CWTMGFP	48249614	46390791(96.15%)	44420291(92.06%)	197
CWTSATGFP	50658270	46331028(91.46%)	43513336(85.9%)	281
CWTVATGFP	29296032	28162889(96.13%)	27249272(93.01%)	91

Showing 1 to 8 of 8 entries Previous 1 Next

- **sample**: SampleID.
- **total_reads**: Total clean reads used for analysis.
- **total_map**: Number and percentage of reads aligned to the genome, the ratio should higher than 70%, total mapping rate: (mapped reads)/(total reads)*100.
- **unique_map**: Number and percentage of reads aligned to the unique position of the reference genome (for subsequent quantitative data analysis), unique mapping rate: (uniquely mapped reads)/(total reads)*100.
- **Multi_map**: number and percentage of reads aligned to multiple locations in the reference genome, multiple mapping rate: (multiple mapped reads)/(total reads)*100.
- **read1_map**: Number and percentage of read1 aligned to the reference genome.
- **read2_map**: Number and percentage of read2 aligned to the reference genome.
- **positive_map**: Number and percentage of reads aligned to the positive chain of the reference genome.
- **negative_map**: Number and percentage of reads aligned to the negative chain of the reference genome.
- **splice_map**: Number of spliced reads on the genome and its percentage.
- **unsplice_map**: Number of complete reads aligned to genome and its percentage.
- **proper_map**: Number of paired read1 and read2 aligned to the genome and its percentage.

2.2.2 Reads Distribution in Reference Genome

Mapped regions can be classified as exons, introns, or intergenic regions. Exon-mapped reads should be the most abundant type of reads when the reference genome is well-annotated. Intron-reads may be derived from pre-mRNA contamination or intron-retention from alternative splicing. Reads mapped to intergenic regions are mainly attributed to weak annotations of the reference genome. The distribution of sequencing reads of all samples in the genomic region is shown in the figure below. See the results file: Mapping/1.Region (./result_tree.html).

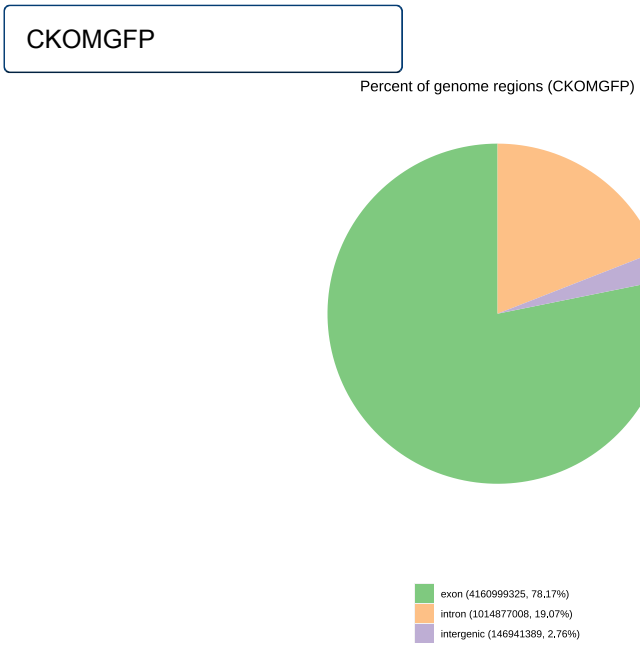


Figure 2.5 Sequencing reads in the genomic region

Note: The ratios of the different colors in the figure represent the ratio of reads to different regions

- **exon**: The number of reads aligned to exon regions of the genome and its proportion in clean reads.
- **Intron**: The number of reads aligned to intron regions of the genome and its proportion in clean reads.
- **Intergenic**: The number of reads aligned to intergenic regions of the genome and its proportion in clean reads.

2.2.3 Visualization of Mapping Results

Files provided in BAM format--a standard file format that contains mapping results–indicating the information of the corresponding referenced genome and gene annotations for some species. The Integrative Genomics Viewer (IGV) is a recommended software for visualizing data from BAM files.

IGV can be featured as below:

1. Displaying the positions of single or multiple reads in the reference genome, and read distribution between annotated exons, introns or intergenic regions, both in adjustable scale respectively;
2. Displaying the read abundance of different regions to demonstrate their expression levels, in adjustable scale;
3. Providing annotation information for both genes and splicing isoforms;
4. Providing other related annotation information;
5. Displaying annotations downloaded from remote servers and/or imported local machines.
6. IGV browser usage can refer to our provided documentation IGVQuickStart (src/IGVQuickStart.pdf)

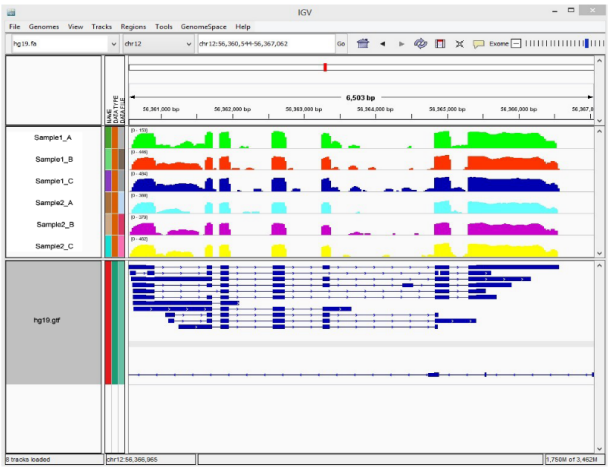


Figure 2.6 IGV Browser Comparison Visualization of Results

2.2.4 Alignment Q&A

What is the difference between the RNA data comparison software and the DNA data comparison software?

Due to the existence of alternative splicing during transcription, a large portion of the measured reads spans different exons, and all RNA data comparison software supports reads' splice comparison.

What are the reasons for the lower rate of mapping?

- The assembly of reference genome is not satisfying

- The relative relation between the tested species and the reference genome is far
- Special treatments or exogenous contaminations of samples

Does the mapping use full length of reads, or does the mapping use reads that are processed at the beginning and the end?

In our standard procedure, we use the standard RNA-seq kit whose indexes are in the middle of the adapters and the sequencing will be executed for the sequences with indexes. In this way, the sequences of read 1 and read 2 are from the samples and there is no need to process the sequences in mapping. Even if the adapters are in sequences or the quality is low, the read will be removed.

Novogene Co.,Ltd.

2.3 Gene Expression Level Analysis

Gene expression level analysis is the core task in the RNA-seq experiment. Gene expression level is calculated by the number of mapped reads.

Novogene Co.,Ltd.

2.3.1 Gene Expression Quantification

The abundance of transcripts reflects gene expression level directly. In RNA-seq experiments, gene expression level is estimated by the abundance of transcripts (count of sequencing) that mapped to genome or exon. Read counts is proportional to gene expression level, gene length and sequencing depth. FPKM (short for the expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced) is the most common method of estimating gene expression levels, which takes the effects into consideration of both sequencing depth and gene length on counting of fragments (Mortazavi et al., 2008). The data are summarized in the table below. See result file: Quant/1.Count/gene_count.xls (./result_tree.html).

Table 2.4 Quantitative results of gene expression

gene_id	HKOSATGFP	HWTSATGFP	CKOMGFP

ENSMUSG00000064351	142772	80406	128937
ENSMUSG00000037742	124991	63819	99123
ENSMUSG00000097971	77404	604151	99829
ENSMUSG00000024610	42450	25409	29186
ENSMUSG00000034994	53629	23755	39247

Showing 1 to 5 of 5 entries Previous 1 Next

- **gene_id**: Gene number.
- **sample**: Raw read count values of each sample.
- **gene_name**: Gene name.
- **gene_chr**: The name of the chromosome where the gene is located.
- **gene_start**: The starting position of the chromosome where the gene is located.
- **gene_end**: The end position of the chromosome where the gene is located.
- **gene_strand**: The positive and negative strand information of the chromosome where the gene is located.
- **gene_length**: Gene length, the sum of genes from the beginning to the end of all non-overlapping exon regions in chromosomes.
- **gene_biotype**: Gene type,such as coding protein genes,long non-coding genes,etc.
- **gene_description**: Gene description.
- **gene_tf_family**: Transcription factor family annotation information of a gene.

2.3.2 Distribution of Gene Expression Levels

To compare gene expression levels under different conditions, the distribution of gene expression levels and FPKM[8] among different samples are displayed by boxplots. For biological replicates, the final FPKM will be the mean value. Result Directory: Quant/3.Distribution/boxplot.svg (./result_tree.html).

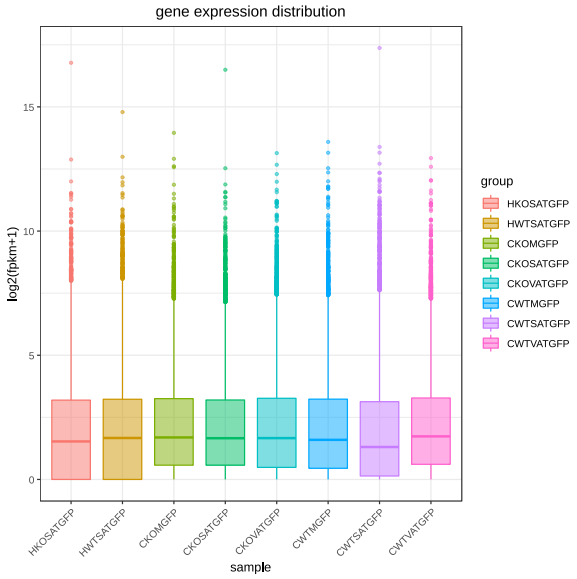


Figure 2.7 Sample gene expression distribution box plot

X axis represents the name of the sample. Y axis indicates the log2(FPKM+1). Parameters of box plots are indicated, including maximum, upper quartile, mid-value, lower quartile and minimum.

2.3.3 Correlation Analysis

Biological replicates are necessary for any biological experiment including RNA-seq technology. Correlation of the gene expression levels between samples plays an important role in verifying reliability and sample selection, which can not only demonstrate the repeatability of the experiment but estimate the differential gene expression analysis.

The closer the correlation coefficient is to 1, the higher similarity the samples have. Encode suggests that the square of the Pearson correlation coefficient should be greater than 0.92 (under ideal experiment conditions) and the R^2 should be greater than 0.8.

According to all gene expression levels (RPKM or FPKM) of each sample, the correlation coefficient of samples between groups is calculated and drawn as heat maps. It is intuitive to show sample differences and repeat cases between groups. The higher the correlation coefficient of the sample is, the closer the expression pattern is. The correlation coefficient matrix is shown in the following figure. Result Directory: Quant/2.Correlation/correlation.svg (./result_tree.html).

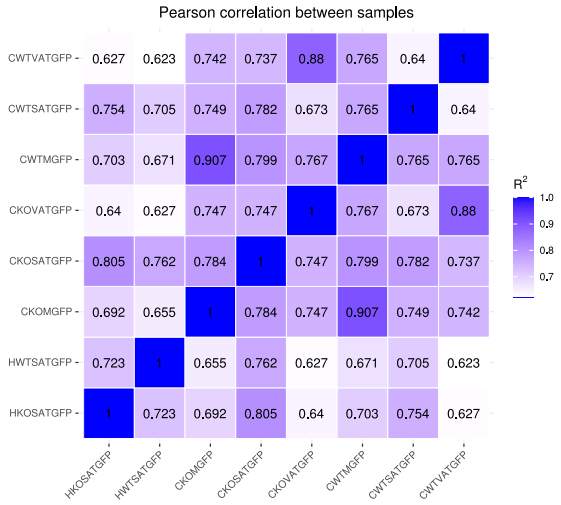


Figure 2.8 Inter-sample correlation heat map

R^2: Square of Pearson correlation coefficient(R)

2.3.4 Principal Component Analysis

Principal component analysis (PCA) is also commonly used to evaluate intergroup differences and intragroup sample duplication. PCA uses the linear algebra calculation method to reduce dimension and extract principal components from tens of thousands of gene variables. We performed PCA analysis on the gene expression value (FPKM) of all samples, as shown in the figure below. Under ideal conditions, the samples between groups should be dispersed and the samples within groups should be gathered together. See the result file: Quant/2.Correlation/pca.svg (./result_tree.html).

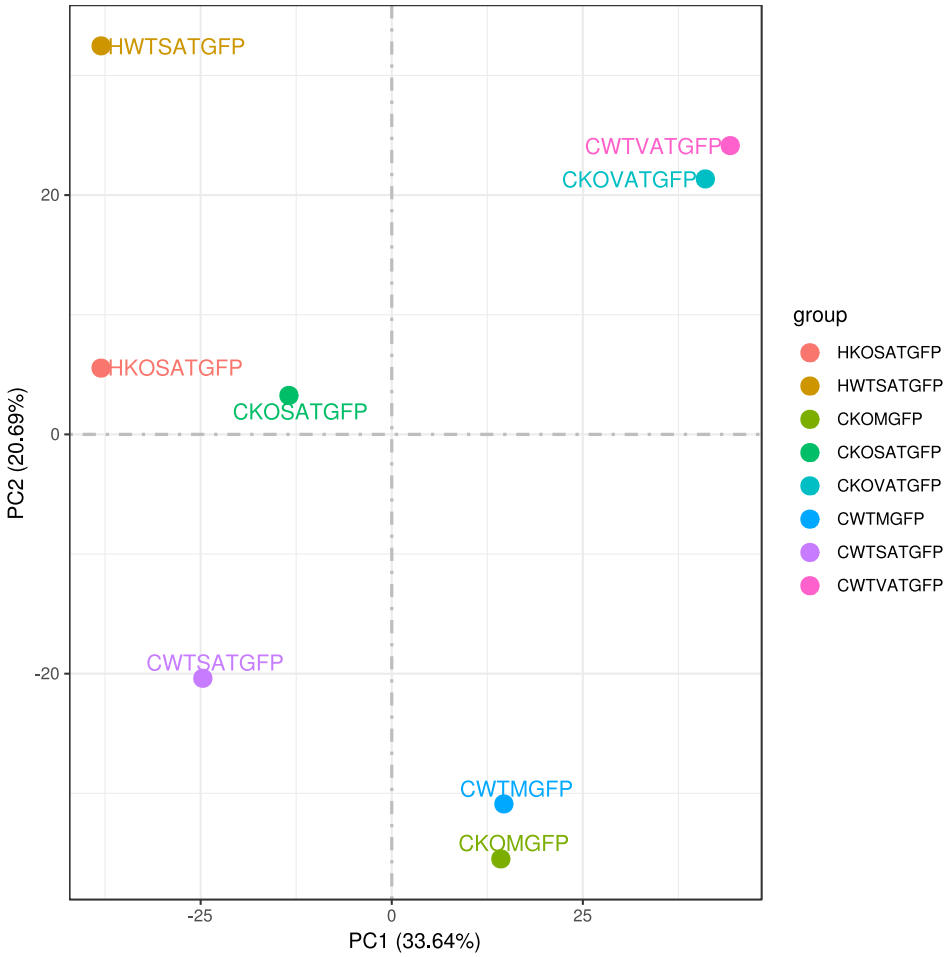
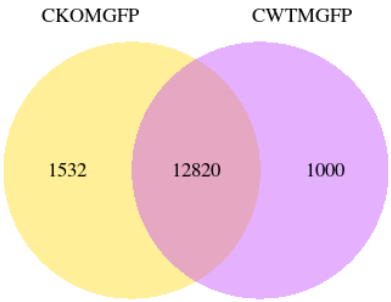


Figure 2.9 Principal component analysis result

2.3.5 Coexpression Venn Diagram

The coexpression Venn diagram presents the number of genes that are uniquely expressed within each group/sample, with the overlapping regions showing the number of genes that are co-expressed in two or more groups/samples. Result Directory: Quant/4.coExpression_venn (./result_tree.html)

venn1



2.3.6 Quantification Analysis Q&A

How do you deal with reads (reads multi mapped) mapped to multiple locations in the genome during the quantification analysis?

Reads mapped to multiple locations in the genome are not possible to be determined their corresponding genes. Therefore, these reads are directly filtered out during the quantification analysis.

For two genes with overlapping regions, how are the reads in the overlapping region assigned during the quantification analysis?

Same as the previous answer.

How is the information of tf_family annotation of a gene obtained in the quantification analysis table?

On the one hand, it is annotated by the tf database (AnimalTFDB/PlantTFDB) (for the species already included in the database). On the other hand, it is predicted by protein domain databases such as Pfam/SUPERFAMILY.

How is FPKM calculated?

The calculation of FPKM(expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced) takes into account the influence of sequencing depth and gene length on the count of fragments, and it is a commonly used method for estimating gene expression levels (Trapnell, Cole, et al., 2010).

What is the threshold for gene expression levels? Why is this threshold set?

Gene expression is generally considered to be greater than 1 for FPKM, a threshold recommended by mainstream journals.

What is the significance of the correlation between samples? How is it calculated?

The sample correlation represents the similarity among samples. The sample correlation could help us examine the similarity at the gene expression level. Higher correlation means higher similarity and less number of differential express genes. Generally, the correlation between biological replicates should be higher than the correlation of samples with different sources. There are three different calculation methods: A. Pearson correlation; B. Spearman rank correlation; C. Kendall's τ. Novogene uses R language to calculate the Pearson correlation coefficient.

What is the pca?

Principal Component Analysis (PCA) is a multivariate dimensionality reduction analysis method. The key of PCA is to reduce the dimension of the data under the premise of keeping the data as much as possible. Briefly, This method will ignore the less relevant variables to describe the relations between different samples. A dataset is a group of points in a multidimensional space and PCA method will move all the points to a new coordinate system without changing their relative spatial position, making their projections have the largest variances in the new coordinate. In this new coordinate system, the axis with the largest variance of projection is PC1 and the second largest one is PC2.

2.4 Differential Gene Expression Analysis

After the gene expression is quantified, statistical analysis of the expression data is required to screen the genes whose expression levels are significantly different in different conditions. The differential analysis is mainly divided into three steps.

- First, the raw readcount is normalized, mainly to correct the sequencing depth;
- Next, the statistical model is used to calculate the hypothesis test's probability (pvalue);
- Finally, multiple hypothesis test corrections are used to obtain FDR values (false discovery rate).

For different experimental conditions, we selected appropriate software for gene expression differential analysis, as shown in the following table.

Table 2.5 Software for Differential Analysis and Differential Gene Screening Criteria

Type	Software	Normalized Method	pvalue Calculation Model	FDR Calculation Method	Differential Gene Screening Threshold
Biological Replicates	DESeq2(Anders et al., 2014)	DESeq	The Negative Binomial Distribution	BH	|log2(FoldChange)| >= 1 & padj <= 0.05
No Biological Replicates	edgeR(Robinson et al., 2010)	TMM	The Negative Binomial Distribution	BH	|log2(FoldChange)| >= 1 & padj <= 0.05

If the number of different genes screened according to the above threshold is too small (less than 100), there will likely be no significant results in the subsequent functional enrichment analysis. Therefore, we will appropriately reduce the threshold for screening different genes according to the specific conditions of the project. If the project experiment only focuses on the expression of a few genes (such as gene knockout), please ignore the enrichment results, and filter the genes of interest from the differential analysis table below.

In general, if a gene differs more than twice as much in expression in both sets of samples, we believe that such genes are differentially expressed. In order to judge whether the difference in expression between two samples is due to various errors or essential differences, we need to make a hypothesis test on the expression data of all genes in these two samples. The transcriptome analysis is performed on thousands of genes, which leads to the accumulation of false positives. The more the number of genes, the higher the cumulative degree of false positives in the hypothesis test, so the introduction of padj to the hypothesis test P-value is calibrated to control the proportion of false positives [13].

The screening criteria for differential genes are very important. The standard we give |log2(FoldChange)| >= 1 & padj<= 0.05 is a commonly used empirical value in actual projects. It can be flexibly selected according to the situation. For example, the

difference multiple can be 1.5 times, or 3 times, padj include 0.01, 0.05, 0.1, etc. If the number of differential genes screened according to the above criteria is too small, it is likely that there will be no significant results in the subsequent functional enrichment analysis. If the project experiment only focuses on the expression of a few genes (such as gene knockout), please ignore the enrichment results, and filter the genes of interest from the differential analysis table below. Conversely, if the number of differential genes obtained is too large, which is not conducive to the screening of subsequent target genes, screening can be performed using stricter threshold criteria at this time, and screening can be performed using stricter threshold criteria.

2.4.1 Result of Differential Expression Analysis

The differential significance analysis for each compare group is shown in the table below. The table shows the first 30 rows of the differential significance results for all the first compare group. See the results file: Differential/1.deglist (./result_tree.html).

Table 2.6 Differential Gene List Partial Results Presentation

gene_id	HKOSATGFP	HWTSATGFP	log2
ENSMUSG00000064337	4742.86550893809	13.8067055631991	8.41
ENSMUSG00000064339	10126.7044755728	259.028029641858	5.28
ENSMUSG00000097300	1.38777878078145e-17	243.873962097464	-11.
ENSMUSG00000076258	56882.7000452457	2303.44819174676	4.62
ENSMUSG00000094797	1.38777878078145e-17	210.810537102697	-10.7

Showing 1 to 5 of 5 entries Previous 1 Next

- **gene_id**: Gene id.
- **sample**: The readcount values of each sample after normalization.
- **group**: The standardized average of readcount for each group of samples.
- **log2FoldChange**: The ratio of gene expression level between the treatment group and the control group was processed by the shrinkage model of the differential analysis software, and finally the logarithm was taken with 2 as the base.
- **pvalue**: Pvalue in hypergenometric tests.
- **padj**: The corrected pvalue of multiple hypothesis test.
- **gene_name**：Gene name.
- **gene_chr**: The name of the chromosome where the gene is located.
- **gene_start**: The starting position of the chromosome where the gene is located.
- **gene_end**: The end position of the chromosome where the gene is located.
- **gene_strand**: The positive and negative strand information of the chromosome where the gene is located.
- **gene_length**: Gene length, the sum of genes from the beginning to the end of

all non-overlapping exon regions in chromosomes.

- **gene_biotype**: Gene type,such as coding protein genes,long non-coding genes,etc.
- **gene_description**: Gene description.
- **gene_tf_family**: Transcription factor family annotation information of gene.

The significance analysis result of each gene in all compare groups was shown in the following table. See the results file: 0.SupFile/all_compare.xls (./result_tree.html).

Table 2.7 Differences in the significance of each gene in all comparison combinations Show

gene_id	HKOSATGFP_count	HWTSATGFP_count
ENSMUSG00000064351	142772	80406
ENSMUSG00000037742	124991	63819
ENSMUSG00000097971	77404	604151
ENSMUSG00000024610	42450	25409
ENSMUSG00000034994	53629	23755

Showing 1 to 5 of 5 entries Previous 1 Next

- **gene_id**: Gene number.
- **sample_count**: Raw readcount value of each sample.
- **sample_fpkm**: FPKM value of each sample.
- **compare_treat**: The average readcount value of treatment group in compare group after normalized.
- **compare_control**: The average readcount value of control group in compare group after normalized.
- **compare_log2FoldChange**: The ratio of gene expression level between the treatment group and the control group of a comparison combination was processed by the shrinkage model of the differential analysis software, and finally the logarithm was taken with 2 as the base.
- **compare_pvalue**: Pvalue in hypergenometric test of compare group.
- **compare_padj**: The corrected pvalue of multiple hypothesis test of compare group.
- **gene_name**: gene name.
- **gene_chr**: The name of the chromosome where the gene is located.
- **gene_start**: The starting position of the chromosome where the gene is located.
- **gene_end**: The end position of the chromosome where the gene is located.
- **gene_strand**: The positive and negative strand information of the chromosome where the gene is located.

- **gene_length**: Gene length,the sum of genes from the beginning to the end of all non-overlapping exon regions in chromosome.
- **gene_biotype**: Gene type,such as coding protein genes,long non-coding genes,etc.
- **gene_description**: Gene description.
- **gene_tf_family**: Transcription factor family annotation information of gene.

2.4.2 Differential Gene Statistics

The statistics of the number of differential genes (including up-regulation and down-regulation) for each compare group and the threshold for screening are shown in the table below, see the results file: Differential/1.deglist/diff_stat.xls (./result_tree.html).

Table 2.8 Differential Gene Statistics Results

compare	all	up	down	
HKOSATGFPvsHWTSATGFP	6398	2969	3429	edgeR padj<=0.05 |lc
CKOSATGFPvsCWTSATGFP	8209	5248	2961	edgeR padj<=0.05 |lc
CKOVATGFPvsCWTVATGFP	1156	520	636	edgeR padj<=0.05 |lc
CKOMGFPvsCWTMGFP	1224	767	457	edgeR padj<=0.05 |lc

Showing 1 to 4 of 4 entries Previous 1 Next

- **compare**: Compare group name.
- **all**: The total number of differential genes in the compare group.
- **up**:The up-regulation number of differential genes in the compare group.
- **down**: The down-regulation number of differential genes in the compare group.
- **threshold**: The compare group software and thresholds for differential gene screening.

The number of differential genes (including up-regulation and down-regulation) for each comparison combination is shown in a histogram:

diff_stat

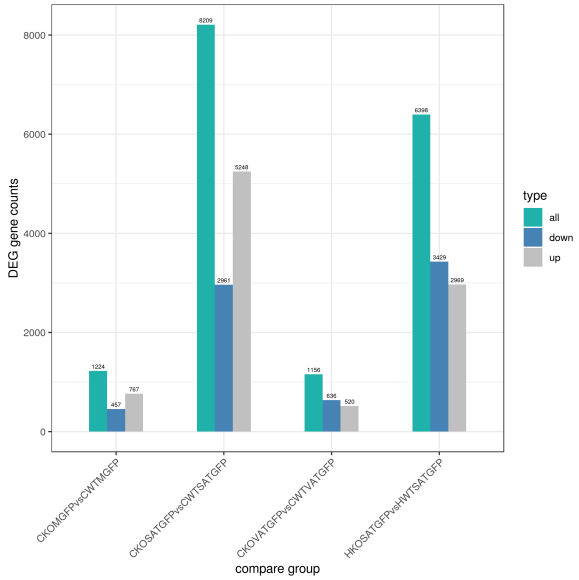


Figure 2.10 Difference Comparison Combine Differential Gene Number Statistics Histogram

Note: Blue and gray represent the differential genes for down-regulation and up-regulation, respectively, and the numbers on the columns indicate the number of differential genes

Volcano plots can be used to infer the overall distribution of differentially expressed genes. In the figure, The x-axis shows the fold change in gene expression between different samples, and the y-axis shows the statistical significance of the differences. Red dots represent up-regulation genes and green dots represent down-regulation genes. See the result file: Differential/1.deglist/{comparative combination}/_volcano.png (./result_tree.html).

CKOMGFPvsCWTMGFP

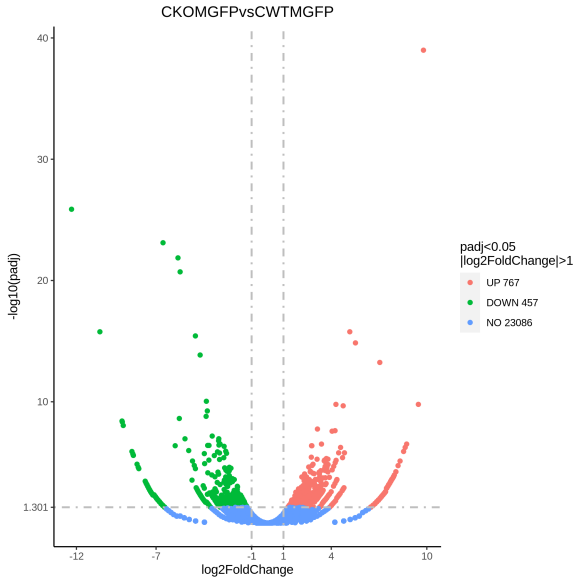


Figure 2.11 Differential Gene Volcano Map

The abscissa in the figure is log2FoldChange, and the ordinate is -log10padj or -log10pvalue, the blue dashed line indicates the threshold line for differential gene screening criteria

2.4.3 Cluster Analysis

All the differentially expressed genes in the comparison group were pooled as the differential gene set. For more than two groups of experiments, cluster analysis can be carried out on different gene sets and genes with similar expression patterns can be clustered together. We used the mainstream hierarchical clustering to cluster the fpkm values of genes, and homogenized the row (Z-score). The genes or samples with similar expression patterns in the heat map will be gathered together. The color in each grid reflects not the gene expression value, but the value obtained after homogenizing the expression data rows (generally between - 2 and 2). Therefore, the colors in the heat map can only be compared horizontally (the expression of the same gene in different samples), but not vertically (the same sample). There are not only inter group clustering, but also inter sample clustering. The final report shows the clustering among samples. See the result file: Differential/3.cluster/heatmap.png (./result_tree.html).

heatmap

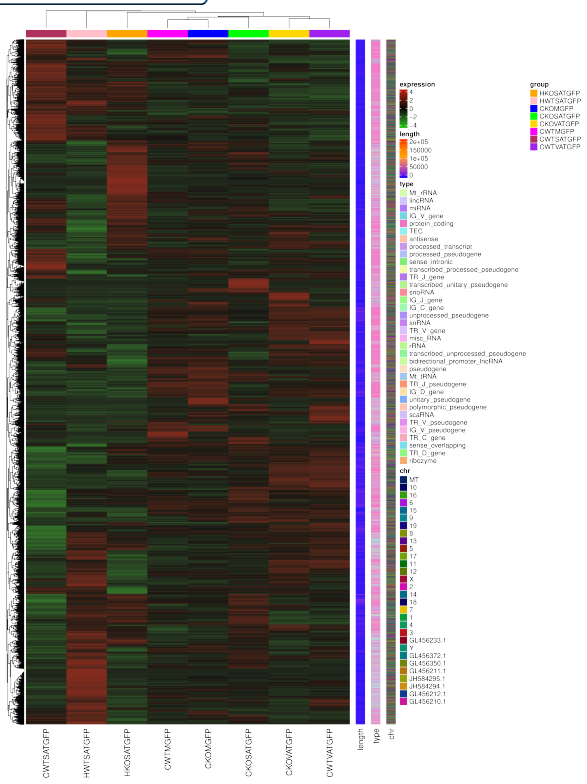


Figure 2.12 differential expression gene clustering heatmap

The overall results of FPKM cluster analysis, clustered using the log2(FPKM+1) value. Red color indicates genes with high expression levels, and green color indicates genes with low expression levels. The color ranging from red to green indicates that log2(FPKM+1) values wherefrom large to small. The chromosome to which each gene belongs, gene's length, and the biological type of the gene are also added to the heatmap

2.4.4 Differential Gene Expression Analysis Q&A

Can I use FPKM for differential analysis?

Most of the differential analysis software (DESeq, DESeq2 and edgeR) use the raw read counts as the input file and the negative binomial distribution model to estimate the probability of gene differential expression between samples. The software itself does some corrections on the read count (mainly the depth of sequencing), while FPKM is the corrected expression value, so it is unreasonable to use FPKM for the differential analysis to make double corrections.

Why do multiple hypothesis tests calculate padj values instead of using pvalues directly to screen fordifferential genes?

There is no problem with the usage of pvalue for a single hypothesis test, but in the process of differential analysis, we have to perform a hypothesis test on each gene. There are often tens of thousands of genes in a species, and tens of thousands of hypothesis tests are performed, which leads to greatly increasing false positives. Assuming a pvalue of 0.05 (only five of the one hundred differential genes are false positives), this accuracy is sufficient for a gene that is hypothesized, but for the entire tens of thousands of genes, such accuracy is far from good. For example, for every 10,000 genes tested, around 500 genes are false positive. In order to properly control false positive rates, it is necessary to introduce a stricter indicator which is the corrected pvalue. Of course, if there are too few differential genes, you can also use pvalues instead. As long as it has biological significance and can be verified by experiments.

What is the maximum threshold that can be set for differential gene screening? Is it appropriate to lower the threshold properly?

In general, the presetting screening threshold in higher-level article is more stringent, and in some articles, the differential gene screening thresholds are appropriately lower. As in some articles without biological replicates, only padj is used as a differential gene screening threshold, regardless of log2foldchange. While some other articles use pvalue as a screening threshold for differential genes.

What are the steps in cluster analysis?

Logarithmic transformation of gene expression values (FPKM expression matrices) of all samples is performed to approximate the data to an ideal normal distribution (clustering analysis rely on an early assumption that the data is subject to an idealized statistical distribution, usually Normal Distribution), then the software calculates the distance (Euclidean distance) of all the transformed data points, and finally classifies N objects into k groups by the complete clustering method in hierarchical clustering. The objects in each group are mutually similar. The hierarchical clustering algorithm is mainly divided into three steps, as shown below.

- At the initial moment, all points are themselves a cluster.
- Find the nearest two clusters to form a cluster. The distance between the two clusters refers to the distance between the two nearest points in the cluster.
- Repeat the second step until all points are clustered into individual cluster.

How is normalization calculated in cluster analysis?

The values of each row of data in the expression matrix are subtracted from the mean of each row of data, and divided by the standard deviation of each row of data. The range of values after normalizationare generally in the interval [-2,2].

How is the transcription factor predicted?

Identification of animal transcription factors uses the animal transcription factor database(animalTFDB 3.0), and identification of plant transcription factors uses the plant transcription factor database(PlantTFDB PlantTFDB 4.0). For the species included in the database, if Ensembl geneid is used, the transcription factor is directly screened; the Ensembl geneid is subjected to SUPERFAMILY and Pfam annotation by interproscan software, and the ID of SUPERFAMILY and Pfam of each gene are obtained. Then, the annotations information of the transcription factor family corresponding to each SUPERFAMILY and Pfam in the DBD (Transcription factor prediction database) are used to predict transcription factors. The fungal transcription factor is predicted by the SUPERFAMILY and Pfam annotations by interproscan software, and the ID of SUPERFAMILY and Pfam of each gene are obtained, and then each of the already annotated databases in the

DBD(Transcription factor prediction database) are utilized. The information of the transcription factor family corresponding to SUPERFAMILY and Pfam is predicted.

2.5 Functional Analysis

Through the enrichment analysis of the differential expressed genes, we can find out which biological functions or pathways are significantly associated with differentially expressed genes. Novogene uses the clusterProfiler (Yu G, 2012) software for enrichment analysis, including GO Enrichment, DO Enrichment, KEGG and Reactome database Enrichment etc,.

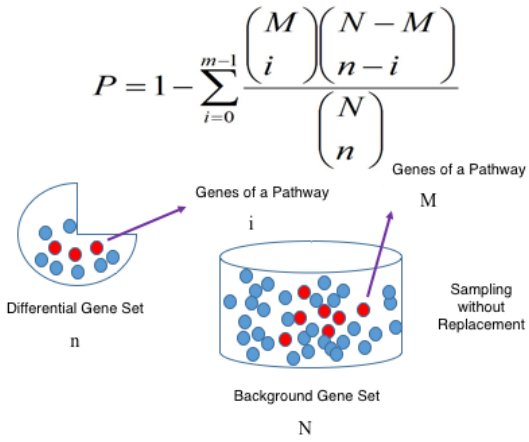


Figure 2.13 Gene Enrichment Analysis Schematic

2.5.1 GO Enrichment Analysis

GO is the abbreviation of Gene Ontology (http://www.geneontology.org/ (http://www.geneontology.org/)), which is a major bioinformatics classification system to unify the presentation of gene properties across all species. It includes three main branches: cellular component, molecular function and biological process. GO terms with padj < 0.05 are significant enrichment. See the result file: Enrichment/GO (./result_tree.html).

Table 2.9 Partially displayed results of GO enrichment analysis of differential genes

Category	GOID	Description	G
BP	GO:0007186	G-protein coupled receptor signaling pathway	1
BP	GO:0007600	sensory perception	1
BP	GO:0006814	sodium ion transport	
BP	GO:0007586	digestion	
BP	GO:0009582	detection of abiotic stimulus	

Showing 1 to 5 of 5 entries Previous 1 Next

- **Category**: Classification of GO databases, including biological processes(BP), cellular components(CC), molecular function(MF).
- **GOID**: Unique identification id of Gene Ontology database.
- **Description**: Function description corresponding to the GO number.
- **GeneRatio**: Ratio between the number of differentially expressed genes in each GO term and all differentially expressed genes that can be found in GO database.
- **BgRatio**: In background GO database, the ratio of all genes concerning this GO term to all genes.
- **pvalue**: Statistics category term; abbreviation for probability value.
- **padj**: Adjusted p-value. Generally, GO Terms with Corrected_pValue < 0.05 are significant enrichment.
- **geneID**: Differentially expressed genes in this term.
- **geneName**: Differentially expressed genes in this term.
- **Count**: The number of difference gene annotated to GO number.
- **Up**: Number of up expressed genes concerning this GO term.
- **Up_Gene_id**: Up expressed gene id concerning this GO term.
- **Down**: Number of down expressed genes concerning this GO term.
- **Down_Gene_id**: down expressed gene id concerning this GO term.

In the results of the GO enrichment analysis, the most significant 30 Terms were selected for display. If the results are less than 30, all terms would be drawn, as shown in the following figure. In this figure, the abscissa is GO Term. The ordinate is the significance level of GO Term enrichment. Higher values correspond to higher significance. The different colors represent the three GO subclasses of BP，CC，MF. See the result file: Enrichment/GO (./result_tree.html).

CKOMGFPvsCWTMGFP.all

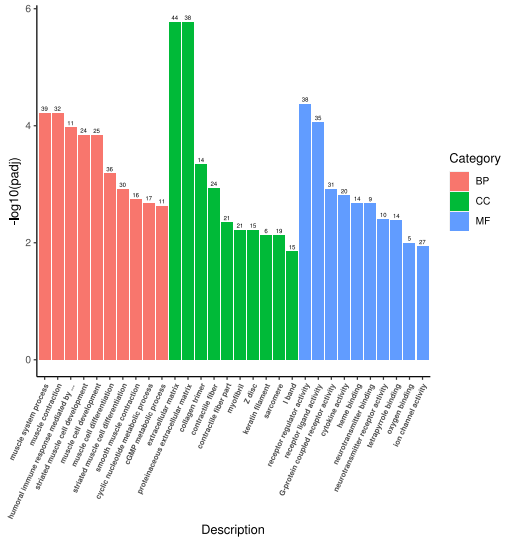
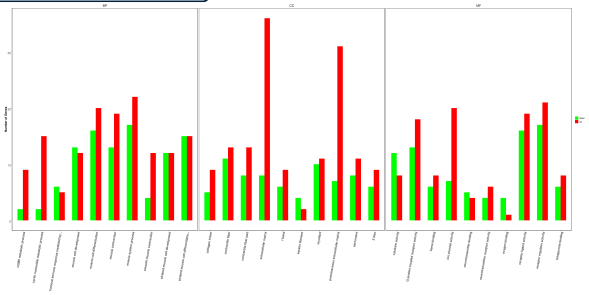


Figure 2.14 GO enrichment analysis histogram

Note: The abscissa in the figure is GO Term, and the ordinate is GO Term's level of significance of enrichment, expressed as -log10(padj). Different colors represent different functional categories

From the GO enrichment analysis result, the most significant 30 Terms were selected for display. If the results are less than 30, all Terms would be drawn according to major categories of biological processes, cell components, molecular functions and categories of up and down expressed genes.

CKOMGFPvsCWTMGFP.all_



The abscissa in the figure is GO Term, and the ordinate is GO Term The level of significance of the set

From the GO enrichment analysis results, the most significant 30 GO Terms were selected for display. If the results are less than 30, all Terms would be drawn. In the figure, the abscissa is the ratio of the number of differential genes linked with the GO Term to the total number of differential genes, and the ordinate is GO Term. The size of a point represents the number of genes annotated to a specific GO Term, and the color from red to purple represents the significant level of the enrichment. See the result file: Enrichment/GO (./result_tree.html).

CKOMGFPvsCWTMGFP.all

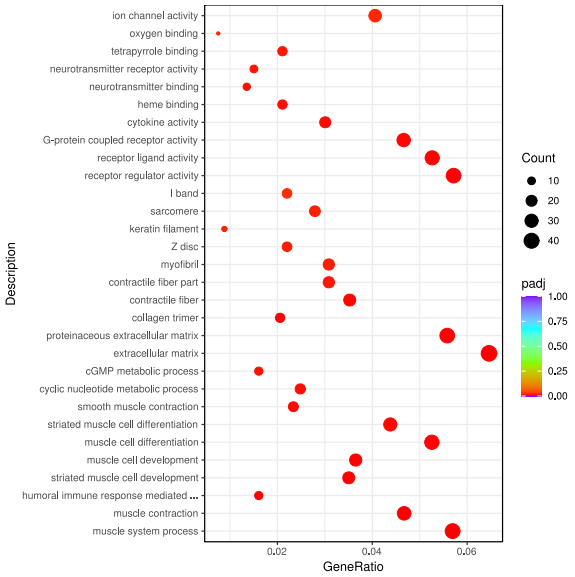


Figure 2.15 GO Enrichment Analysis Scatter Plot

The abscissa in the graph is the ratio of the differential gene number to the total number of differential genes on the GO Term, and the ordinate is GO Term

(Directed Acyclic Graph， DAG) could visualize the enriched GO Term of differential expression genes and its hierarchy. In this graph, branch means the hierarchical relation, and the function ranges become increasingly specific from the top to bottom. In general, the top 5 results of GO enrichment analysis are chosen as the main nodes (shown by box) in directed acyclic graph, and related GO Term are shown together by hierarchical connections. The enrichment degree is illustrated by color shades, the darker the shades are, the higher the enrichment degree is. In this project, DAG of biological process, molecular function and cellular component were drawn sequentially. See the result file: Enrichment/GO (./result_tree.html).

CKOMGFPvsCWTMGFP.all.(

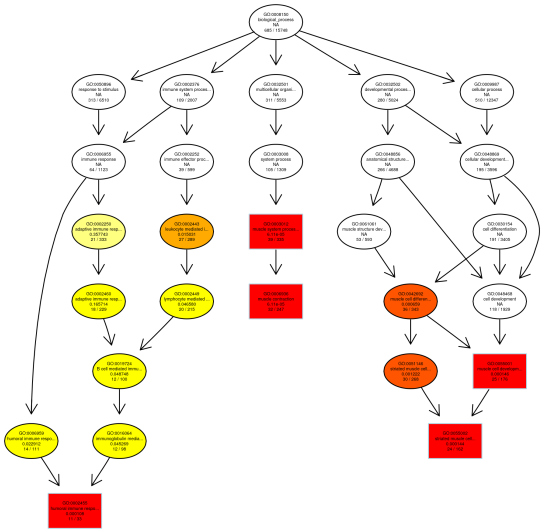


Figure 2.16 GO Enrichment Analysis DAG Chart

Each node represents a GO term, and the box represents the enrichment level of TOP5 GOTerms. The depth of the color represents the degree of enrichment, the darker the color is, the higherthe enrichment degree is. Each node shows the name of the term and the Padj of enrichment analysis

2.5.2 KEGG Enrichment Analysis

The interactions of multiple genes may be involved in certain biological functions. KEGG (Kyoto Encyclopedia of Genes and Genomes) is a collection of manually curated databases containing resources on genomic, biological-pathway and disease information (Kanehisa,2008). Pathway enrichment analysis identifies significantly enriched metabolic pathways or signal transduction pathways associated with differentially expressed genes, comparing the whole genome background. KEGG pathways with padj < 0.05 are significant enrichment. See the result file: Enrichment/KEGG (./result_tree.html).

Table 2.10 KEGG Enrichment Analysis Partial Results

KEGGID	Description	GeneRatio	BgRatio
mmu04080	Neuroactive ligand-receptor interaction	47/1173	102/5342
mmu00380	Tryptophan metabolism	18/1173	32/5342
mmu00590	Arachidonic acid metabolism	18/1173	36/5342
mmu00982	Drug metabolism - cytochrome P450	16/1173	31/5342
mmu00591	Linoleic acid metabolism	8/1173	12/5342

Showing 1 to 5 of 5 entries Previous 1 Next

- **KEGGID**: Unique identification id of KEGG database.
- **Description**: Function description of this pathway.
- **GeneRatio**: Ratio between the number of differentially expressed genes in each pathway and all differentially expressed genes that can be found in KEGG database.
- **BgRatio**: In background KEGG database, the ratio of all genes concerning this KEGG pathway to all genes included.
- **pvalue**: Statistics category term; abbreviation for probability value.
- **padj**: Adjusted p-value. Generally, KEGG pathways with Corrected_pValue < 0.05 are significant enrichment.
- **geneID**: Differentially expressed genes in this pathway.
- **geneName**: Differentially expressed genes in this pathway.
- **keggID**: Pathway gene id.
- **Count**: Number of differentially expressed genes concerning this pathway.

- **Up**: Number of up expressed genes concerning this KEGG pathway.
- **Up_Gene_id**: Up expressed gene id concerning this KEGG pathway.
- **Down**: Number of down expressed genes concerning this KEGG pathway.
- **Down_Gene_id**: Down expressed gene id concerning this KEGG pathway.

In the KEGG enrichment results, the most significant 20 KEGG pathways were selected for display. If the results are less than 20, all pathways would be drawn, as shown in the following figure. In this figure, the abscissa is the KEGG pathway, and the ordinate is the significance level of the pathway enrichment. Higher values correspond to higher significance. See the result file: Enrichment/KEGG (./result_tree.html).

CKOMGFPvsCWTMGFP.all

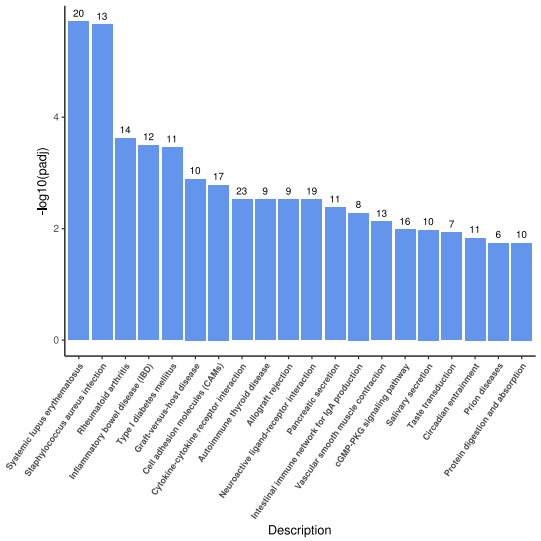


Figure 2.17 KEGG enrichment analysis histogram

The abscissa is the KEGG pathway, and the ordinate is the significance level of pathway enrichment

From the KEGG enrichment results, the most significant 20 KEGG pathways were selected for display. If the results are less than 20, all pathways would be drawn. In this figure, the abscissa is the ratio of the number of differential genes linked with the KEGG pathway to the total number of differential genes. The ordinate is KEGG Pathway. The size of a point represents the number of genes annotated to a specific KEGG pathway. The color from red to purple represents the significant level of the enrichment. See the results file: Enrichment/KEGG (./result_tree.html).

CKOMGFPvsCWTMGFP.all

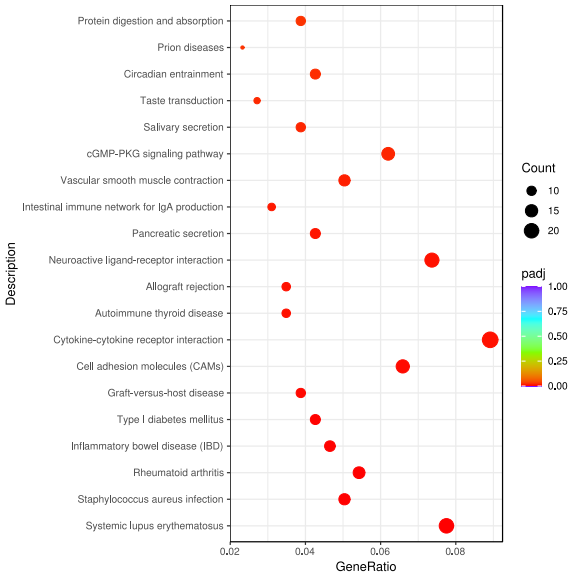


Figure 2.18 KEGG enrichment scatter plot

The abscissa in the graph is the ratio of the number of differential genes on the KEGG pathway to the total number of differential genes, and the ordinate is KEGG pathway

The html file can interactively help you view the significantly enriched KEGG pathway map by clicking corresponding links. In the map, the KEGG node including up-regulated genes is marked red, and the KEGG node including down-regulated genes is marked green. While yellow means the node contains both types of genes. Hovering over the marked KEGG node will show details of differentially expressed genes, with the same color as above, and the number inside brackets is log2(Foldchange). The above steps can be implemented offline. If you are connected to the Internet, click on each node to get the specific information of each KEGG node in the KEGG database. See the result file: Enrichment/KEGG (./result_tree.html).

HKOSATGFPvsHWTSATGFF

The most enriched pathway terms

Statistic method: hypergeometric test

FDR correction method: Benjamini and Hochberg

Term	Description	Sample number	Background number	

2.5.3 Reactome Enrichment Analysis

The Reactome database brings together the various reactions and biological pathways of human model species. Reactome pathway enrichment with padj less than 0.05 as the threshold for significant enrichment, the enrichment results are shown in the following table. See the results file: Enrichment/Reactome (./result_tree.html).

Table 2.11 Reactome Enrichment Analysis Partial Results

ReactomeID	Description	GeneRatio	BgRa

R-MMU-373076	Class A/1 (Rhodopsin-like receptors)	63/1588	139/67
R-MMU-500792	GPCR ligand binding	72/1588	172/67
R-MMU-388396	GPCR downstream signalling	140/1588	415/67
R-MMU-372790	Signaling by GPCR	143/1588	430/67
R-MMU-397014	Muscle contraction	47/1588	111/67

Showing 1 to 5 of 5 entries Previous 1 Next

- **ReactomeID**: Unique identification id of Reactome database.
- **Description**: Function description of this pathway.
- **GeneRatio**: Ratio between the number of differentially expressed genes in each Reactome ID and all differentially expressed genes that can be found in Reactome database.
- **BgRatio**: In background Reactome database, the ratio of all genes annotated on this Reactome pathway and all genes.
- **pvalue**: Statistics category term; abbreviation for probability value.
- **padj**: Adjusted p-value. Generally, Reactome pathways with Corrected_pValue < 0.05 are significant enrichment.
- **geneID**: Differentially expressed genes related to this pathway.
- **geneName**: Differentially expressed genes in this pathway.
- **EntrezID**: Entrez gene ID of differentially expressed genes.
- **Count**: Number of differentially expressed genes concerning this pathway.
- **Up**: Number of up expressed genes concerning this Reactome pathway.
- **Up_Gene_id**: Up expressed gene id concerning this Reactome pathway.
- **Down**: Number of down expressed genes concerning this Reactome pathway.
- **Down_Gene_id**: Down expressed gene id concerning this Reactome term.

Novogene **Novogene Co.,Ltd.**

In the Reactome enrichment results, the most significant 20 Reactome pathways were selected for display. If the results are less than 20, all pathways would be drawn, as shown in the following figure. In the figure, the abscissa is the Reactome pathway, and the ordinate is the significance level of the pathway enrichment. Higher values correspond to high significance. See the result file: Enrichment/Reactome (./result_tree.html).

CKOMGFPvsCWTMGFP.all

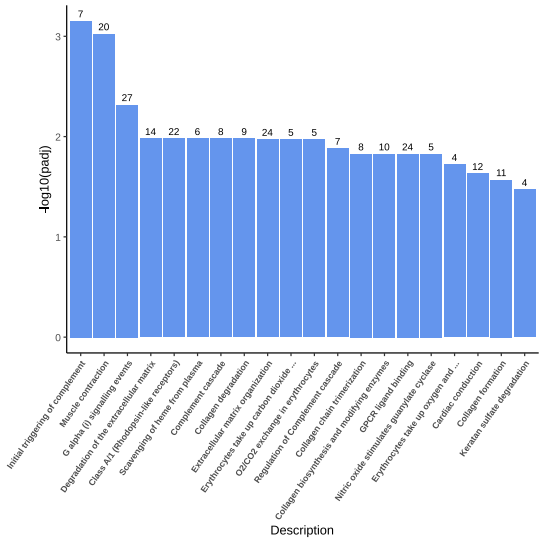


Figure 2.19 Reactome Enrichment Analysis Histogram

The abscissa is the Reactome pathway, and the ordinate is the significance level of pathway enrichment

From the Reactome enrichment analysis results, the most significant 20 Reactome pathways were selected for display. If the results are less than 20, all pathways would be drawn. In the figure, the abscissa is the ratio of the number of differential genes linked with the Reactome pathway to the total number of differential genes, and the ordinate is Reactome Pathway. The size of a point represents the number of genes annotated to a specific Reactome pathway, and the color from red to purple represents the significant size of the enrichment. See the result file: Enrichment/Reactome (./result_tree.html)

CKOMGFPvsCWTMGFP.all

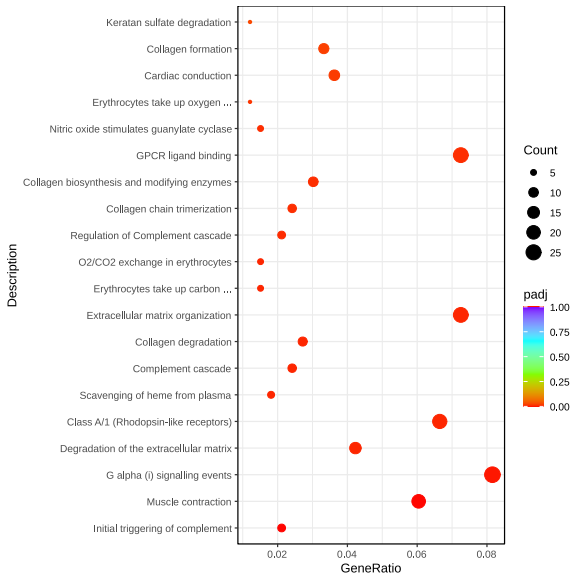


Figure 2.20 Reactome enrichment analysis scatter plot

The abscissa in the graph is the ratio of the number of differential genes on the Refectome pathway to the total number of differential genes in the Reactome pathway. The ordinate is Reactome pathway. See results file: Enrichment/Reactome (./result_tree.html)

2.5.4 Protein-Protein Interaction Network Analysis

The protein-protein interaction network is constructed for differential expression gene by searching STRING protein interaction database STRING (http://string-db.org/), the result file can be found in Enrichment/PPI (./result_tree.html).

Protein-protein interaction is provided as a network file that can be imported into Cytoscape software for visualization and edition. The central organizing metaphor of Cytoscape is a network graph, with molecular species represent as nodes and intermolecular interactions represent as edges.

1. Customized network data display using powerful visual styles.
2. View superposition of gene expression ratios and p-values on the network. Expression data can be mapped to node color, label, border thickness, or border color, etc. according to user-configured colors and visualization schemes.
3. Layout networks in two dimensions. A variety of layout algorithms are available, including cyclic and spring-embedded layouts.
4. Zoom in/out and pan for browsing the network.
5. Use the network manager to easily organize multiple networks. And this structure can be saved in a session file.
6. Use the bird's eye view to easily navigate large networks.
7. Easily navigate large networks (100,000+ nodes and edges) by an efficient rendering engine.

Novogene Co.,Ltd.

2.5.5 Enrichment Analysis Q&A

What is the meaning of n, N, i, M in the enrichment analysis schematic?

In the enrichment analysis results, the column of GeneRatio represents the differential gene set information. The number to the left of the slash represents the number of differential genes annotated to a pathway that corresponds to i of the enrichment schematic. The number to the right of the slash represents the differential gene annotated to all pathways that corresponds to n of the enrichment analysis schematic. The column of

BgRatio represents the background gene set information. The number to the left of the slash is the number of background genes annotated to the pathway, corresponding to the M of the enrichment analysis schematic, and to the right of the slash is annotated background genes in all pathways, corresponding to the N of the enrichment analysis schematic.

Does the enrichment analysis result only look at the channels of significant enrichment?

The interpretation of the results of enrichment analysis should be based on biological significance. pvalue and padj only provide a reference. As long as the results can be explained and have biological significance, those pathways that are not significantly enriched are worthy of further study.

3 Appendix

3.1 Result file decompression method

Compressed.format	Customer.type	method
compressed files in the format of *.tar.gz:	Unix/Linux/Mac user	use tar -zxvf *.tar.gz command
	Windows user	use uncompressed software such as WinRAR, 7-Zip et al
compressed files in the format of *.gz:	Unix/Linux/Mac user	use gzip –d *.gz command
	Windows user	use uncompressed software such as WinRAR, 7-Zip et al
compressed files in the format of *.zip:	Unix/Linux/Mac user	use unzip *.zip command
	Windows user	use uncompressed software such as WinRAR, 7-Zip et al

3.2 Result file format description

File.type	Document.description	Open.mode

file.fa/fasta	sequence file, in the format of fasta in genaral. Since sequence of gene or genome is large, we provide customer with sample.*fasta files(partial sequences of* .fasta). It will be more convenient to check file fomat.	unix/Linux/Mac users use less or more commands to view sequences in the format of *.fasta.
		windows users use editor Editplus/Notepad++ et al
file.fq/fastq	eads sequence file, in the format of fasta. it is not easy to open since it is a large big file.	unix/Linux/Mac users use less or more commands;
		windows users use editor Editplus/Notepad++ et al
file.txt/xls	table result file; files are separated by(Tab)	unix/Linux/Mac users use less or more commands
		windows users use editor Editplus/Notepad++ et al, also can use Microsoft Excel to open.
file.pdf/svg	Results image files and vector images can be enlarged and reduced without distortion, which is convenient for users to view and edit. Adobe Illustrator can be used to edit images and publish articles	Windows / Mac users can use adobe reader / foxin reader / Web browser to open it
		Unix / Linux users use evince command to open
file.png	Results image file; bitmap, lossless compression	Unix / Linux / Mac users use the display command to open
		windows users use picture viewer to view, for example Photoshop etc. al

3.3 Analysis software list and version

Analysis	Software	Version	Parameter	Remarks
Mapping	hisat2	2.0.5	Default	Mapping to a reference
Quantification	featureCounts	1.5.0-p3		
Differential Analysis	DESeq2	1.20.0	|log2(FoldChange)| >= 1 & padj <= 0.05	For sample with bio-replicate
	edgeR	3.22.5	|log2(FoldChange)| >= 1 & padj <= 0.05	For sample without bio-replicate
Enrichment Analysis	clusterProfiler	3.8.1	padj < 0.05	For GO, KEGG enrichment analysis
GSEA Analysis	gsea	v3.0	Default	
Protein-Protein Interaction Analysis	diamond	0.9.14	e-value = 1e-10	Using blast, String database.
Alternative splicing	rMATS	4.1.0	Default	
SNP/InDel Analysis	GATK	4.1.4.1	MQ < 40.0 and QD < 2.0 and FS > 30.0 and DP < 10 and QUAL < 20	SNP/InDel calling
	snpEff	4.3q	Default	SNP/InDel Annotation

3.4 methods

In order to facilitate users to write articles, we have prepared data analysis related toEnglishmethods (src/methods.pdf)

3.5 references

- [1] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics[J]. Nature Reviews Genetics, 2009, 10(1): 57-63.
- [2] Parkhomchuk D, Borodina T, Amstislavskiy V, et al. Transcriptome analysis by strand-specific sequencing of complementary DNA[J]. Nucleic acids research, 2009, 37(18): e123-e123.
- [3] Goldstein L D , Cao Y , Pau G , et al. Prediction and Quantification of Splice Events from RNA-Seq Data.[J]. Plos One, 2016, 11(5):e0156132.
- [4] Mortazavi A, Williams B A, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA-Seq[J]. Nature methods, 2008, 5(7): 621-628.
- [5] Mihaela Pertea, Geo M Pertea, Corina M Antonescu1,et al.StringTie enables improved reconstruction of a transcriptome from RNA-seq reads[J].Nat Biotechnol. 2015 March ; 33(3):290-295(StringTie)
- [6] Liao Y1, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomicfeatures.Bioinformatics.2014 ,30(7):923-30.(featureCounts)
- [7] Garber M, Grabherr M G, Guttman M, et al. Computational methods for transcriptome annotation and quantification using RNA-seq[J]. Nature methods, 2011, 8(6): 469-477.
- [8] Bray N, Pimentel H, Melsted P, et al. Near-optimal RNA-Seq quantification[J]. arXiv preprint arXiv:1505.02710, 2015.
- [9] Patro R, Mount S M, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms[J]. Nature biotechnology, 2014, 32(5): 462-464.
- [10] Anders S, Huber W. Differential expression analysis for sequence count data[J]. Genome biol, 2010, 11(10): R106.
- [11] Love M I, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2[J]. Genome biology, 2014, 15(12): 1-21.(DESeq2)
- [12] Robinson M D, McCarthy D J, Smyth G K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data[J]. Bioinformatics, 2010, 26(1): 139-140.(edgeR)
- [13] Young M D, Wakefield M J, Smyth G K, et al. Method Gene ontology analysis for RNA-seq: accounting for selection bias[J]. Genome Biol, 2010, 11: R14.
- [14] He Z , Zhao X , Lu Z , et al. Comparative transcriptome and gene co-expression network analysis reveal genes and signaling pathways adaptively responsive to varied adverse stresses in the insect fungal pathogen, Beauveria bassiana[J]. Journal of Invertebrate Pathology, 2017:S0022201117304391.
- [15] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes[J]. Nucleic acids research, 2000, 28(1): 27-30.(KEGG)
- [16] Shen S., Park JW., Lu ZX., Lin L., Henry MD., Wu YN., Zhou Q., Xing Y. rMATS: Robust and Flexible Detection of Differential Alternative Splicing from Replicate RNA-Seq Data.(rMATS).
- [17] Katz Y, Wang E T, Airoldi E M, et al. Analysis and design of RNA sequencing experiments for identifying isoform regulation[J]. Nature methods,

2010, 7(12): 1009-1015.
- [18] McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data[J]. Genome research, 2010, 20(9): 1297-1303.(GATK)